AD________________

Award Number: W81XWH-11-1-0713

TITLE: Identification of Novel, Inherited Genetic Markers for Aggressive PCa in European and African Americans Using Whole Genome Sequencing

PRINCIPAL INVESTIGATOR: Jielin Sun, PhD

CONTRACTING ORGANIZATION:

Wake Forest University Health Sciences
Winston Salem, NC 27157

REPORT DATE: September 2013

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT:

■ Approved for public release; distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
September 2013

2. REPORT TYPE
ANNUAL

3. DATES COVERED *(From - To)*
22 August 2012 – 21 August 2013

4. TITLE AND SUBTITLE
Identification of Novel, Inherited Genetic Markers for Aggressive PCa in European and African Americans Using Whole Genome Sequencing

5a. CONTRACT NUMBER

5b. GRANT NUMBER W81XWH-11-1-0713

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Jielin Sun; Siqun Lilly Zheng

email: jisun@wakehealth.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Wake Forest University Health Sciences
Medical Center Boulevard
Winston Salem, NC 27157

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and
Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Prostate cancer (PCa) is the most common cancer and the second leading cause of cancer death among men in the United States. While most prostate cancer patients have an indolent form of the disease that may not even require treatment, about 10-15% of PCa patients have an aggressive form that may progress to metastases and death, thus requiring intensive treatment. Several clinical variables such as PSA levels, Gleason grade and TNM stage are good predictors for disease with poor clinical outcomes; however, their predictive performance needs to be improved. Our inability to reliably distinguish between these two forms of PCa, early on in the course of the disease has resulted in the over-treatment of many and under treatment of some. The identification of additional markers, including genetic variants will improve our ability to distinguish aggressive from indolent forms of PCa and to better understand the racial disparity of PCa that exists between Europen Americans (EA) and African Americans (AA). In this DOD proposal, we hypothesized that multiple rare sequence variants in the genome may increase aggressive PCa risk. Through a genome-wide search of rare variants based on an existing population from Johns Hopkins Hospital (JHH) of 600 aggressive PCa patients and 600 indolent PCa patients using Illumina's Human Exome BeadChip, we identified several rare variants that are significantly associated with aggressive PCa development in EA or AA populations. The implicated rare variants will be followed in additional populations.

15. SUBJECT TERMS
Prostate cancer, indolent, lethal (aggressive), sequence variants

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**c. THIS PAGE**	**19a. NAME OF RESPONSIBLE PERSON USAMRMC**
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	16	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

INTRODUCTION

While most prostate cancer (PCa) patients have an indolent form of the disease that may not even require treatment, about 10-15% of PCa patients have an aggressive form that may progress to metastases and death, thus requiring intensive treatment. Several clinical variables such as PSA levels, Gleason grade, and TNM stage are good predictors for disease with poor clinical outcomes; however, their predictive performance needs to be improved. Our inability to reliably distinguish between these two forms of PCa, early on in the course of the disease has resulted in the over-treatment of many and under treatment of some. Another dilemma is a large difference in PCa risk, especially aggressive PCa, between races. African Americans (AAs) have the world's highest incidence of PCa and are twice as likely, as compared with Caucasians to die of the disease. Inherited markers of aggressive PCa could be used for screening and diagnosis of aggressive PCa at an early stage while reducing over-diagnosis and treatment for others. The overall hypothesis is that inherited sequence variants in the genome are associated with a lethal (aggressive) form of PCa but not indolent PCa, and the difference in these variants between races may contribute to higher incidence of and mortality from aggressive PCa in AA.

In this DOD proposal, we proposed: 1) To discover novel inherited genetic variants in the genome that may be associated with aggressive but not indolent PCa using an exome array approach; **2)** To confirm the novel genetic variants using mass spectrometry directed sequencing; and **3)** To perform association tests of implicated genetic variants among 1,500 most aggressive PCa and 1,500 least aggressive (i.e. indolent) PCa.

BODY

Approved Revised Statement of Work:

Aim 1. To discover novel inherited genetic variants in the genome that may be associated with aggressive but not indolent PCa using a WGS approach.

Step by Step method and expected results

1. **Months 1-6:** Preparation of the study, including regulatory review, IRB approval and other logistical issues
2. **Months 7-12**: Perform exome SNP array analysis for 400 (200 aggressive PCa and 200 indolent PCa) cases in EAs and 400 (200 aggressive PCa and 200 indolent PCa) cases in AAs from Johns Hopkins Hospital.
3. **Months 13-18:** Perform exome SNP array analysis for 200 (100 aggressive PCa and 100 indolent PCa) cases in EAs and 200 (100 aggressive PCa and 100 indolent PCa) cases in AAs from Johns Hopkins Hospital. Perform statistical and bioinformatics analysis for the combined dataset of 600 aggressive PCa cases and 600 indolent PCa cases.

Outcome and deliverables

We expect to identify a certain number of novel rare variants most likely associated with aggressive but not indolent PCa.

Aim 2. To confirm the genetic variants implicated in Aim 1 using Sequenom

Step by Step method and expected results

1. **Months 19-22:** Genotyping the top rare mutations among the additional PCa samples using Sequenom
2. **Months 23-24**: Confirmation analysis of the top SNPs

Outcome and deliverable

We expect that a subset of the top rare mutations will be confirmed using the Sequenom platform.

Aim 3. To perform association tests of selected genetic variants among 1,500 most aggressive PCa and 1,500 most indolent PCa.

Step by Step method and expected results

1. **Months 25-26:** Genotyping ~100 SNPs in 1,500 most aggressive PCa and 1,500 most indolent PCa patients
2. **Months 27-28:** Perform association test of these SNPs with aggressiveness of PCa using a logistic regression model
3. **Months 29-36:** Final analysis and preparation of papers

Outcome and deliverable
We expect to identify several novel rare mutations that are associated with aggressiveness of PCa using exome SNP array analysis. We will prepare and submit papers reporting the major results from the study.

Summary report

By Sep 2013, we were in the 24th month of this funded project. During the last year, we have completed the following 1) performed genotyping of exome-array among additional 200 aggressive PCa and 200 indolent PCa in European American (EA) and AA (African American) samples; 2) performed single rare variant analysis, bioinformatics analysis, as well as gene-based analysis (SKAT) in the combined dataset of 600 aggressive PCa and 600 indolent PCa samples to identify rare variants that have strong effects on PCa risk; 3) confirmed the top genetic variants implicated in Aim 1 using Sequenom platform.

Detailed report

Study Subjects. Subjects included in the John Hopkins (JHH) study were recruited during Jan. 1999 to Dec. 2008. All of them underwent radical prostatectomy for treatment of prostate cancer. Details of this study have been described in previous publications. In this study, aggressive prostate cancer was defined as: 1) Gleason Score ≥8; or 2) Gleason Score =7, with the most prevalent pattern being 4; or 3) stage T3b or higher; or 4) involvement of regional lymph nodes; or 5) presence of distant metastasis. Otherwise, the cancers were classified as non-aggressive prostate cancer. In this study, a total of 1,200 subjects (including 600 EA and 600 AA samples) from JHH study were genotyped using the Illumina Human Exome BeadChip platform.

Genotyping and Quality Control. Genotyping was conducted using the Illumina Human Exome BeadChip at the Center for Cancer Genomics, Wake Forest University School of Medicine. A total of 247,870 genetic variants were included in the ExomeArray. Among them, 92,173 and 102,366 genetic variants were polymorphic in EA, and AA samples, respectively. Those polymorphic SNPs were used for sex and IBS check of all subjects using PLINK software (Purcell 2007). In addition, polymorphic SNPs were also used to estimate the missing rate per individual.

For polymorphic genetic variants, only those with a missing rate >0.98 in subjects passed QC and were kept for further statistical analyses. Thus, a total of 91,998 and 98,644 variants in EA and AA samples were included in further analyses.

Bioinformatics analysis (Variant effect prediction): All coding nonsynonymous variants were assessed for potential effect by Polymorphism Phenotyping version 2 (PolyPhen2), which is a tool for predicting the possible impact of an amino acid substitution on the structure and function of a human protein. For a given variant, PolyPhen2 calculates a Naïve Bayes posterior probability that the mutation is damaging and then appraised qualitatively as benign, possibly damaging, or probably damaging (Adzhubei 2010).

Statistical analysis for single SNP effect. Principal components analysis was conducted to detect potential population stratification by EIGENSTRAT software (Price 2006). The top 5 eigenvectors which indicates ancestral heterogeneity within a group of individuals were adjusted as covariates in multivariate logistic regression analysis.

All polymorphic genetic variants that passed QC were evaluated for associations with prostate cancer aggressiveness. For genetic variants with any of the genotype counts ≤5, Fisher's exact test was applied to investigate potential association. For genetic variants with genotype counts >5, multivariate logistic regression analysis was conducted assuming an additive genetic model, adjusting for age-at-diagnosis and the top 5 eigenvectors. All analyses were performed using the PLINK software package (Purcell 2007).

Gene-based analysis: We used a novel statistical approach called Sequence Kernel Association Test (SKAT), to conduct gene-based analysis of rare variants for aggressive PCa. SKAT is a supervised and flexible regression method to test for association between rare variants in a gene or genetic region and a continuous or dichotomous trait. Compared to other methods of estimating the joint effect of a subset of SNPs, SKAT is able to deal with variants that have different direction and magnitude of effects, and allows for covariate adjustment (Wu 2011). In addition, SKAT can also avoid arbitrary selection of threshold in burden test. Moreover, SKAT is computationally efficient, compared to a permutation test, making it feasible to analyze the large dataset in our study.

Results

EA population

Single SNP analysis

The top SNPs that were significantly associated with aggressive PCa based on 300 aggressive PCa cases and 300 indolent PCa cases in EAs are listed in Table 1-2. A total of 13 SNPs were identified with low frequency (MAF < 0.05) at a P-value cutoff of 1E-03 (Table 1). Those 13 SNPs were located on 12 genomic regions. The top significant SNP, rs78649652, was located on the *ERAP1* gene on chromosome 5, with a MAF of 0.04 in aggressive PCa and 0.006 in indolent PCa (P = 9.59E-05). Men who carry the "A" allele had 6.54 fold increased risk for aggressive PCa, compared with men carrying the "G" allele (OR = 6.54).

Table 1. Top signficant variants with low frequency (MAF<0.05) associated with aggressive PCa in EAs from JHH population

SNP	CHR	BP	A1	A2	Maf_Agg	Maf_NonAgg	OR	P_Fisher	Category	Function	GeneName
rs78649652	5	96,124,373	A	G	0.040	0.006	6.543	9.59E-05	nonsynonymous	missense	*ERAP1*
rs79954845	17	36,483,889	C	G	0.029	0.003	9.439	2.42E-04	nonsynonymous	missense	*GPR179*
rs17101661	14	64,564,680	A	G	0.048	0.011	4.431	2.66E-04	nonsynonymous	missense	*SYNE2*
rs34075341	9	91,616,843	A	G	0.004	0.045	0.079	2.82E-04	nonsynonymous	missense	*S1PR3*
rs61818256	1	201,294,910	A	G	0.059	0.018	3.483	3.67E-04	nonsynonymous	missense	*PKP1*
bs19_9068458	19	9,068,458	A	T	0.026	0.002	10.980	3.99E-04	nonsynonymous	missense	*MUC16*
rs16950981	18	6,992,683	A	T	0.037	0.007	5.263	5.76E-04	nonsynonymous	missense	*LAMA1*
bs2_242593011	2	242,593,011	A	G	0.033	0.006	6.077	5.82E-04	nonsynonymous	missense	*ATG4B*
rs56224008	9	131,107,634	A	G	0.066	0.023	2.984	7.09E-04	nonsynonymous	missense	*SLC27A4*
rs61729839	2	238,277,379	A	G	0.044	0.011	4.075	7.50E-04	nonsynonymous	missense	*COL6A3*
rs16830693	1	43,805,240	G	A	0.063	0.021	3.138	7.62E-04	Splice	silent	*MPL*
rs17851681	1	227,954,677	A	G	0.048	0.113	0.395	7.97E-04	nonsynonymous	missense	*SNAP47*

rs17851681	1	227,954,677	A	G	0.048	0.113	0.395	7.97E-04	nonsynonymous	missense	*SNAP47*

A total of 36 SNPs were identified with relatively common frequency (MAF > 0.05) at a P-value cutoff of 1E-03 (Table 2). Those 36 SNPs were located on 21 genomic regions. The top 14 significant SNPs were located on *FSIP2* gene on chromosome 2. The allele “A” of the top SNP, rs17228441, was present more frequent in the aggressive PCa (60.3%), compared with indolent PCa (44.7%), with a P-value of 3.48E-6. Men who carry the “A” allele had 1.88 fold increased risk for aggressive PCa, compared with men carrying the “G” allele (OR = 1.88).

Table 2. Top signficant variants with common frequency(MAF>0.05) associated with aggressive PCa in EAs from JHH population

SNP	CHR	BP	A1	A2	Maf Agg	Maf NonAgg	OR	P_Fisher	Category	Function	GeneName
rs17228441	2	186,627,943	A	G	0.603	0.447	1.878	3.48E-06	Nonsynonymous	missense	*FSIP2*
rs992822	2	186,654,592	A	G	0.603	0.448	1.871	3.62E-06	Nonsynonymous	missense	*FSIP2*
rs17229201	2	186,656,956	A	G	0.603	0.448	1.871	3.62E-06	Nonsynonymous	missense	*FSIP2*
rs10490391	2	186,658,438	G	A	0.603	0.448	1.871	3.62E-06	Nonsynonymous	missense	*FSIP2*
rs10490392	2	186,658,565	C	A	0.603	0.448	1.871	3.62E-06	Nonsynonymous	missense	*FSIP2*
rs2161036	2	186,659,359	C	A	0.603	0.448	1.871	3.62E-06	Nonsynonymous	missense	*FSIP2*
rs10931200	2	186,664,963	C	A	0.603	0.448	1.871	3.62E-06	Nonsynonymous	missense	*FSIP2*
rs36004074	2	186,665,432	G	A	0.603	0.448	1.871	3.62E-06	Nonsynonymous	missense	*FSIP2*
rs11695215	2	186,665,824	A	G	0.603	0.448	1.871	3.62E-06	Nonsynonymous	missense	*FSIP2*
rs7605884	2	186,667,121	A	G	0.603	0.448	1.871	3.62E-06	Nonsynonymous	missense	*FSIP2*
rs17826534	2	186,671,357	G	A	0.603	0.448	1.871	3.62E-06	Nonsynonymous	missense	*FSIP2*
rs60029104	2	186,658,056	G	A	0.610	0.455	1.874	3.62E-06	Nonsynonymous	missense	*FSIP2*
rs1862066	2	186,671,912	G	A	0.599	0.447	1.849	6.59E-06	nonsynonymous&Fingerprint	missense	*FSIP2*
rs16827154	2	186,670,780	T	A	0.603	0.453	1.836	9.10E-06	Nonsynonymous	missense	*FSIP2*
rs6883840	5	40,286,410	A	G	0.129	0.240	0.468	3.48E-05	aimList	Intergenic	*PTGER4*
rs714147	23	150,912,962	A	G	0.169	0.054	3.538	3.52E-05	Nonsynonymous	missense	*CNGA2*
rs2418135	9	113,901,309	A	G	0.382	0.520	0.571	4.26E-05	gridList	Intergenic	*OR2K2*
rs12918952	16	78,420,775	G	A	0.294	0.423	0.568	7.25E-05	Nonsynonymous	missense	*WWOX*
rs11921691	3	113,673,125	A	G	0.375	0.506	0.585	9.86E-05	Nonsynonymous	missense	*ZDHHC23*
rs10057851	5	64,565,261	G	A	0.375	0.502	0.596	1.70E-04	gridList	Intron	*ADAMTS6*
rs877859	3	107,714,075	G	A	0.412	0.297	1.658	3.14E-04	aimList	Intergenic	*CD47*
rs61734605	11	34,916,657	A	G	0.423	0.307	1.652	3.51E-04	Splice	missense	*APIP*
rs2247572	8	73,633,028	A	G	0.206	0.120	1.901	3.89E-04	GWAS	Intron	*KCNB2*
rs10274334	7	47,925,331	G	C	0.500	0.382	1.621	4.49E-04	Nonsynonymous	missense	*PKD1L1*
rs938886	14	20,837,701	G	C	0.140	0.234	0.533	4.71E-04	Nonsynonymous	missense	*TEP1*
rs2122554	5	165,957,086	A	C	0.081	0.031	2.733	4.93E-04	GWAS	Intergenic	*ODZ2*
rs6939340	6	22,140,004	G	A	0.581	0.464	1.601	4.95E-04	GWAS	Intron	*LINC00340*
rs2455512	23	88,889,091	A	C	0.309	0.173	2.139	5.27E-04	aimList	Intergenic	*TGIF2LX*
rs1713449	14	20,841,707	A	G	0.136	0.228	0.533	5.65E-04	Nonsynonymous	missense	*TEP1*
rs12961939	18	6,997,818	C	A	0.191	0.290	0.578	7.12E-04	Nonsynonymous	missense	*LAMA1*
rs3133745	8	96,534,806	A	G	0.206	0.125	1.818	9.45E-04	aimList	Intron	*LOC100616530*
rs11955074	5	178,294,060	A	G	0.184	0.108	1.860	9.75E-04	Nonsynonymous	missense	*ZNF354B*
rs10804178	2	210,849,283	G	A	0.401	0.512	0.637	1.02E-03	gridList	Intron	*UNC80*
rs4712653	6	22,125,964	G	A	0.559	0.452	1.536	1.61E-03	GWAS	Intron	*LINC00340*
rs3095250	6	31,208,340	G	A	0.353	0.437	0.703	1.21E-02	HLA	Intergenic	*HLA-C*

rs3130688	6	31,210,216	G	A	0.353	0.435	0.708	1.46E-02	HLA	Intergenic	*HLA-C*

Gene-based analysis

We performed gene-based analysis using the SKAT approach. The top genes with P-value < 1E-03 are presented in Table 3 and Table 4. We first conducted the SKAT analysis based on all variants. Thirty gene sets were identified (Table 3). The top genes associated with aggressive PCa were *CREB3L1*, *KLF13*, *ROBO4,* and *ZCCHC6*, with a P-value range from 4.17E-05 to 2.55E-06.

Table 3. Top signficant genes associated with aggressive PCa using SKAT approach in EAs from JHH population (based on all variants)

Gene	P.value	N.Marker.All	N.Marker.Test
CREB3L1	2.55E-06	9	9
KLF13	1.43E-05	1	1
ROBO4	2.63E-05	9	9
ZCCHC6	4.17E-05	7	7
RNF208	1.01E-04	2	2
LOC152742	1.24E-04	2	2
TRIM17	1.72E-04	3	3
L3MBTL2	1.78E-04	4	4
SNX10	2.01E-04	2	2
CXorf68	2.01E-04	1	1
ZSCAN23	2.01E-04	1	1
F8	2.01E-04	2	2
FAM45A	2.28E-04	1	1
KRTAP22-1	2.63E-04	2	2
RSG1	2.88E-04	3	3
TMEM177	3.11E-04	7	7
CDH6	3.32E-04	4	4
SPAG7	3.40E-04	2	2
RAB26	3.48E-04	3	3
IL16	3.70E-04	12	12
ZNF829	4.24E-04	3	3
EXOC3L2	4.32E-04	2	2
RIMS3	5.03E-04	3	3
MIR4697	5.60E-04	1	1
ARHGEF10	6.36E-04	12	12
C9orf135	6.41E-04	8	8
MLXIPL	6.65E-04	6	6
PNMA2	6.73E-04	3	3
CCL16	9.79E-04	1	1
AHNAK	9.98E-04	32	32

We then conducted the SKAT analysis based on low frequency variants (MAF < 0.05) only. Similar sets of genes were identified and compared with the results based on all variants (Table 4). The top genes associated with aggressive PCa were *CREB3L1*, *ROBO4* and *ZCCHC6*, with a P-value range from 4.17E-05 to 2.64E-06.

Table 4. Top signficant genes associated with aggressive PCa using SKAT approach in EAs from JHH population (based on variants with low frequency (MAF < 0.05))

Gene	P.value	N.Marker.All	N.Marker.Test
CREB3L1	2.64E-06	5	5
ROBO4	2.63E-05	8	8

ZCCHC6	4.17E-05	5	5
RNF208	1.01E-04	2	2
TRIM17	1.72E-04	3	3
L3MBTL2	1.78E-04	4	4
SNX10	2.01E-04	1	1
ZSCAN23	2.01E-04	1	1
CXorf68	2.01E-04	1	1
F8	2.01E-04	1	1
FAM45A	2.28E-04	1	1
RSG1	2.88E-04	3	3
TMEM177	3.06E-04	3	3
CDH6	3.32E-04	1	1
SPAG7	3.40E-04	2	2
RAB26	3.48E-04	3	3
ARHGEF10	3.52E-04	10	10
IL16	3.63E-04	9	9
ZNF829	4.24E-04	3	3
EXOC3L2	4.32E-04	2	2
RIMS3	5.03E-04	3	3
C9orf135	6.41E-04	7	7
MLXIPL	7.61E-04	4	4
LPHN3	9.67E-04	11	11
CCL16	9.79E-04	1	1
AHNAK	9.98E-04	32	32

AA population

Single SNP analysis

The top significant SNPs that were significantly associated with aggressive PCa based on 300 aggressive PCa cases and 300 indolent PCa cases in AAs are listed in Table 5 and Table 6. A total of 11 SNPs were identified with low frequency (MAF < 0.05) at a P-value cutoff of 1E-03 (Table 5). Those 11 SNPs were located on 11 genomic regions. The top significant SNP, rs28382644, was located on the *POLM* gene on chromosome 7, with a MAF of 0.034 in aggressive PCa and 0.004 in indolent PCa (P = 4.64E-05). Men who carry the "G" allele had 8.93 fold increased risk for aggressive PCa, compared with men carrying the "C" allele (OR = 8.93).

Table 5. Top signficant variants with low frequency (MAF<0.05) associated with aggressive PCa in AAs from JHH population

SNP	CHR	BP	A1	A2	Maf_Agg	Maf_NonAgg	OR	P_Fisher	Category	Function	GeneName
rs28382644	7	44,118,394	G	C	0.034	0.004	8.931	4.64E-05	nonsynonymous	missense	*POLM*
rs10090835	8	130,789,767	A	G	0.011	0.056	0.190	1.44E-04	nonsynonymous	missense	*GSDMC*
rs61898615	11	103,019,260	A	G	0.031	0.005	6.524	3.44E-04	nonsynonymous	missense	*DYNC2H1*
rs41289902	6	112,460,365	A	G	0.028	0.004	7.399	4.09E-04	nonsynonymous	missense	*LAMA4*
rs912969	13	103,867,104	A	G	0.028	0.076	0.347	6.72E-04	GWAS	Intergenic	*SLC10A2*
rs7530895	1	203,260,756	G	A	0.008	0.044	0.186	6.72E-04	aimList	Intergenic	*LOC730227*

rs117476305	12	101,761,753	G	A	0.031	0.006	5.431	7.52E-04	splice	missense	*UTP20*
rs118014343	8	623,906	A	G	0.056	0.019	3.000	7.95E-04	nonsynonymous	missense	*ERICH1*
rs117497357	7	20,768,077	A	T	0.073	0.030	2.534	9.60E-04	splice	missense	*ABCB5*
rs8176345	12	58,158,558	A	G	0.061	0.023	2.755	9.73E-04	synonymous	silent	*CYP27B1*
rs34070230	19	4,844,790	C	G	0.061	0.023	2.755	9.73E-04	nonsynonymous	missense	*PLIN3*

A total of 16 SNPs were identified with relatively common frequency (MAF > 0.05) at a P-value cutoff of 1E-03 (Table 6). Those 16 SNPs were located on 16 genomic regions. The top SNP, rs10841496 was located on *PDE3A* gene on chromosome 12. The allele "A" of the top SNP, rs10841496, was present more frequent in the aggressive PCa (57.8%), compared with indolent PCa (46.2%), with a P-value of 1.79E-04. Men who carry the "A" allele had 1.59 fold increased risk for aggressive PCa, compared with men carrying the "C" allele (OR = 1.59).

Table 6. Top signficant variants with common frequency (MAF>0.05) associated with aggressive PCa in AAs from JHH population

SNP	CHR	BP	A1	A2	Maf_Agg	Maf_NonAgg	OR	P_Fisher	Category	Function	GeneName
rs10841496	12	20,521,654	A	C	0.5782	0.4623	1.595	1.79E-04	GWAS	Intergenic	*PDE3A*
rs10499052	6	109,885,475	A	G	0.3547	0.2515	1.637	2.16E-04	nonsynonymous	missense	*AKD1*
rs1901187	17	38,646,147	G	A	0.3464	0.4584	0.6261	2.52E-04	gridList	Intron	*TNS4*
rs2272994	1	40,923,019	A	G	0.2877	0.1925	1.695	2.55E-04	nonsynonymous	missense	*ZNF643*
rs2924461	5	8,012,069	G	A	0.5279	0.4159	1.571	2.68E-04	gridList	Intergenic	*MTRR*
rs4790404	17	2,886,642	G	A	0.3911	0.5019	0.6372	2.91E-04	gridList	Intron	*RAP1GAP2*
rs7236632	18	55,434,202	G	A	0.2235	0.1393	1.779	3.13E-04	GWAS	Intron	*ATP8B1*
rs669408	1	232,519,150	C	A	0.4887	0.3781	1.572	3.44E-04	GWAS	Intergenic	*SIPA1L2*
rs11116595	12	85,165,879	A	G	0.3771	0.4845	0.644	4.48E-04	gridList	Intergenic	*SLC6A15*
rs9701796	1	19,186,129	C	G	0.2793	0.1896	1.657	5.44E-04	nonsynonymous	missense	*TAS1R2*
rs7008113	8	111,438,655	A	G	0.2709	0.1847	1.64	6.54E-04	aimList	Intergenic	*KCNV1*
rs11881700	19	52,538,428	G	A	0.1872	0.1151	1.77	8.20E-04	synonymous	silent	*ZNF432*
rs7787531	7	129,023,597	G	A	0.1229	0.0648	2.022	9.25E-04	GWAS	Intron	*AHCYL2*
rs28568406	1	158,687,163	A	G	0.4637	0.3646	1.507	1.07E-03	nonsynonymous	missense	*OR6K3*
rs741301	7	36,917,995	G	A	0.4134	0.3211	1.49	1.93E-03	GWAS	Intron	*ELMO1*
rs4919060	10	98,699,136	A	C	0.2207	0.148	1.63	2.26E-03	aimList	Intron	*LCOR*

Gene-based analysis

We performed gene-based analysis using the SKAT approach in the AA population. The top genes with P-values < 1E-03 are presented in Table 7 and Table 8. We first conducted the SKAT analysis based on all variants. A total of 22 genes sets were identified (Table 7). The top genes associated with aggressive PCa were *TEK, CDH2 and BEST* , with P-values that ranged from 7.97E-05 to 1.47E-05.

Table 7. Top signficant genes associated with aggressive PCa using SKAT approach in AAs from JHH population (based on all variants)

Gene	P.value	N.Marker.All	N.Marker.Test

TEK	1.47E-05	9	9
CDH2	5.24E-05	14	14
BEST2	7.97E-05	4	4
LOC100130581	1.45E-04	1	1
OR11L1	2.38E-04	8	8
S100PBP	2.59E-04	4	4
LOC643339	4.25E-04	2	2
INMT	5.07E-04	11	11
LOC148145	5.51E-04	1	1
NRIP3	5.80E-04	3	3
PPARGC1B	6.33E-04	13	13
TIMM44	7.05E-04	8	8
LOC401164	7.18E-04	3	3
RFC1	7.47E-04	2	2
SLC16A5	7.47E-04	2	2
SRSF1	7.71E-04	1	1
MORN3	7.79E-04	4	4
CDH3	7.79E-04	10	10
DDHD1	8.71E-04	2	2
TMEM106C	9.02E-04	5	5
KLK15	9.52E-04	4	4
LINC00284	9.80E-04	1	1

We then conducted the SKAT analysis based on low frequency variants (MAF < 0.05) only. Similar sets of genes were identified, compared with the results based on all variants (Table 8). The top genes associated with aggressive PCa were *TEK, CDH2, and BEST2* genes, with P-value ranged from 7.22E-05 to 1.47E-05.

Table 8. Top signficant genes associated with aggressive PCa using SKAT approach in AAs from JHH population (based on variants with low frequency (MAF < 0.05))

Gene	P.value	N.Marker.All	N.Marker.Test
TEK	1.47E-05	8	8
CDH2	6.95E-05	5	5
BEST2	7.22E-05	3	3
OR11L1	2.01E-04	4	4
S100PBP	2.59E-04	4	4
PPARGC1B	4.16E-04	9	9
INMT	5.01E-04	7	7
NRIP3	5.80E-04	3	3
TIMM44	7.05E-04	8	8
CDH3	7.30E-04	5	5
POU1F1	7.47E-04	2	2
RFC1	7.47E-04	2	2
SLC16A5	7.47E-04	2	2
TMEM106C	7.85E-04	3	3
KLK15	9.52E-04	3	3

Confirmation of the top rare variants using Sequenom. The top rare variants implicated in Table 1 and Table 4 were genotyped in the 600 aggressive and 600 indolent PCa samples using Sequenom Platform. All variants were confirmed to be real genetic variants, instead of genotyping error or calling algorithm error of ExomeArray. This indicates that applying stringent QC criteria minimized the impact of genotyping inaccuracies of rare variants.

Discussion

To our knowledge, our study represents one of the first comprehensive studies to identify rare variants that are associated with aggressive PCa in both EAs and AAs. Our data generated from the first two years had identified potential important rare variants that are associated with aggressive PCa.

We selected the Illumina Human Exome BeadChip (ExomeArray) as our genotyping platform to study rare variants. The ExomeArray chip represents the newest gene chip that delivers unparalleled coverage of putative functional exonic variants. The relatively cheaper cost makes it possible to study larger sample sizes. The Exome Beachip is comprised of >240,000 markers, including >200,000 nonsynonymous SNPs, nonsense mutations, SNPs in splice sites and promoter regions, as well as thousands of GWAS tag markers. Nearly 90% of the SNPs on the exome arrays are rare, with a MAF<5%. In addition, the markers on the Illumina Human Exome BeadChips are selected from over 12,000 individual exome and whole-genome sequences, representing diverse populations, including those of European and African descent. Therefore, it is more efficient and economical to use exome arrays to identify rare variants associated with aggressive PCa, compared with whole genome sequencing.

The top rare SNP (rs78649652) implicated in our study is a nonsynonymous SNP located on the *ERAP1* gene. ERAP1 (Endoplasmic Reticulum Aminopeptidase) encodes protein which is an aminopeptidase involved in trimming HLA class I-binding precursors so that they can be presented on human leukocyte antigen (HLA) class I molecules. ERAP1 is an important component in the antigen processing machinery (APM) and HLA class I molecules which are key determinants of immune recognition of bacteria and virus infected cells (Heemels et al 1995). It is crucial for APM component to process endogenous, tumor-associated proteins and therefore can subsequently lead to cytotoxic T lymphocyte-mediated anti-tumor immune reactions (Mehta 2009). Therefore, variants located on genes that are crucial for APM component may affect normal immune response. Recently, genetic variants on the *ERAP1* gene had also been found to be associated with a variety of immune-related diseases, including ankylosing spondylitis (Szczypiorska 2011), juvenile idiopathic arthritis (Hinks 2011), and psoriasis (Trembath 2010). More importantly, genetic variants located on *ERAP1* gene have been reported to be associated with increased risk for cervical carcinoma (Mehta 2009).

In our study, we found that the rare allele “A” of rs78649652 was present more frequently in aggressive PCa (4%) than and in indolent PCa (0.6%). We speculate the variant allele of rs7849652 may be associated with function or downregulation of ERAP1, which may lead to the preferential loading of nontumor-associated peptides. This may yield a less immunogenic phenotype and promote tumor growth and aggressiveness. Our hypothesis is supported by several recent publications that report a less immunogenic peptide in ERAP1-deficient mice (Falk 2002; Saveanu 2005; Hammer 2006; Yan 2006). In this sense, the nonsynonymous variant rs78649652 may affect ERAP1 function, which leads to decreased trimming of relevant epitopes (Mehta 2009). The less immunogenic peptides may result in immune evasion by tumor cells which in turn accelerate tumor progression. More experimental efforts are needed to pinpoint the biological implication of the rare variants on the *ERAP1* gene and protein function.

The gene-based approach (SKAT) we adopted is a novel statistical approach. SKAT is a supervised and flexible regression method to test for association between rare variants in a gene or genetic region and a continuous or dichotomous trait. Compared to other methods of estimating the joint effect of a subset of SNPs, SKAT is able to deal with variants that have different direction and magnitude of effects, and allows for covariate adjustment (Wu 2011). In addition, SKAT can also avoid arbitrary selection of threshold in burden test. Moreover, SKAT is computationally efficient, compared to a permutation test, making it feasible to analyze the large dataset in our study. Interestingly, several of the top targets identified by SKAT analysis (*CREB3L1* and *KLF13*) encode transcription factors.

Specifically, CREB3L1 (CAMP Responsive Element Binding Protein 3-Like 1) encodes a transcription factor that acts during endoplasmic reticulum stress by activating unfolded protein response target genes. It is specifically involved in ER-stress response in astrocytes in the central nervous system. CREB3L1 may also play a role in gliosis. In vitro, CREB3L1 binds to box-B element, cAMP response element (CRE) and CRE-like sequences, and activates transcription through box-B element but not through CRE. KLF13 (Kruppel-like factor 13) belongs to a family of transcription factors that contain 3 classical zinc finger DNA-binding domains consisting of a zinc atom tetrahedrally coordinated by 2 cysteines and 2 histidines (C2H2 motif). These transcription factors bind to GC-rich sequences and related GT and CACCC boxes (Scohy et al., 2000). The function of KLF13 represses transcription by binding to the BTE site, a GC-rich DNA element, in competition with the activator SP1. It also represses transcription by interacting with the corepressor Sin3A and HDAC1. Previous reports had found overexpression of KLF13 in oral cancer cells (Henson 2010). In addition, genetic variants in the gene encoding Kruppel-like factor 7 are associated with type 2 diabetes (Kanazawa 2005).

Besides all the above findings, we have also carefully calculated the study power based on our modified study design. We have >80% power to detect an OR of 2.0 (3.6) for variants with a MAF of 0.05 (0.01), at an alpha level of 1E-05 (2-sided). Therefore, we have sufficient power to identify novel rare mutations with relatively large effect based on our proposed sample size. We also considered several procedures to control for multiple test correction and SNP selection to be confirmed in additional independent samples. The Bonferroni corrected P-values are 2E-7 (0.05/200,000 variants) and 2E-6 (0.05/20,000 genes), for single variant analysis and gene-based analysis, respectively. However, not all the tests for single variants are independent due to linkage disequilibrium (LD) structure among variants. In addition, previous studies also showed that the true associations do not necessarily reach the stringent Bonferroni corrected P-value cutoffs. Therefore, to balance study power and false positives, rare variants in Aim 1 that meet either of the following criteria with less stringent P-value cutoffs will be selected for replication: 1) variants reach a p-value of 1E-3 in single variant analysis; 2) variants in genes which reach a p-value of 1E-3 in gene-based analysis by SKAT. The adoption of the two-stage study design will further help to remove false positives.

In conclusion, we have identified several targeted rare variants and genes that are associated with aggressive PCa. In year 3, we will follow those top variants and genes identified in Aim 1 and Aim 2 in additional samples from the JHH population. This step will eliminate most of the false positives that were identified due to multiple testing issues. The variants that will be confirmed represent variants that are associated with increased risk for aggressive PCa. The newly identified variants can provide more insight into the etiology of aggressive PCa and provide potential effective targets for therapy of aggressive PCa.

KEY RESEARCH ACCOMPLISHMENTS

1) Completed IRB and other logistic issues
2) Performed genotyping of exome-array among additional 200 aggressive PCa and 200 indolent PCa in European American (EA) and AA (African American) samples
3) Performed single rare variant analysis, bioinformatics analysis, and gene-based analysis (SKAT) to identify rare variants that have strong effects on aggressive PCa risk in the combined population of 600 aggressive PCa and 600 indolent PCa samples
4) Confirmed the top rare variants using Sequenom platform

REPORTABLE OUTCOMES

1) Top rare and common variants and genes in the genome that are significantly associated with aggressive PCa in EAs (Table 1 - Table 4)
2) Top rare and common variants and genes in the genome that are significantly associated with aggressive PCa in AAs (Table 5 - Table 8)

CONCLUSION

1) We have made great progress in achieving the goals described in the approved Statement of Work.
2) We have identified a list of rare and common variants in the genome that are associated with aggressive PCa.
3) The top rare variants have been confirmed to be real rare variants, instead of genotyping artifacts, by using another platform of Sequenom. We will follow those SNPs in an additional study population to confirm the association results in year 3 of the funded study.

REFERENCES

Adzhubei I.A., Schmidt S., Peshkin L., et al. A method and server for predicting damaging missense mutations. Nat Methods. 7, 248-249 (2010).

Akbari MR, Trachtenberg J, Lee J, Tam S, Bristow R, Loblaw A, Narod SA, Nam RK. Association Between Germline HOXB13 G84E Mutation and Risk of Prostate Cancer. J Natl Cancer Inst. 2012 Aug 1;104(16):1260-2. Epub 2012 Jul 9.

Castro E. G.C.L., Olmos D., et al. Correlation of germ-line BRCA2 mutations with aggressive prostate cancer and outcome. ASCO Meet Abstr. 29, 1517 (2011).

Ewing CM, Ray AM, Lange EM, Zuhlke KA, Robbins CM, Tembe WD, Wiley KE, Isaacs SD, Johng D, Wang Y, Bizon C, Yan G, Gielzak M, Partin AW, Shanmugam V, Izatt T, Sinari S, Craig DW, Zheng SL, Walsh PC, Montie JE, Xu J, Carpten JD, Isaacs WB, Cooney KA. Germline mutations in HOXB13 and prostate-cancer risk. N Engl J Med. 2012 Jan 12;366(2):141-9.

Falk K, Rötzschke O. The final cut: how ERAP1 trims MHC ligands to size. Nat Immunol. 2002 Dec;3(12):1121-2.

Fredericks WJ, McGarvey T, Wang H, Lal P, Puthiyaveettil R, Tomaszewski J, Sepulveda J, Labelle E, Weiss JS, Nickerson ML, Kruth HS, Brandt W, Wessjohann LA, Malkowicz SB. The bladder tumor suppressor protein TERE1 (UBIAD1) modulates cell cholesterol: implications for tumor progression. DNA Cell Biol. 2011 Nov;30(11):851-64. Epub 2011 Jul 8.

Gallagher D.J., Gaudet M.M., Pal P., et al. Germline BRCA mutations denote a clinicopathologic subset of prostate cancer. Clin Cancer Res. 16, 2115-2121 (2010).

Hammer GE, Gonzalez F, Champsaur M, Cado D, Shastri N. The aminopeptidase ERAAP shapes the peptide repertoire displayed by major histocompatibility complex class I molecules. Nat Immunol. 2006 Jan;7(1):103-12. Epub 2005 Nov 20.

Heemels MT, Ploegh H. Generation, translocation, and presentation of MHC class I-restricted peptides. Annu Rev Biochem. 1995;64:463-91. Review.

Henson BJ, Gollin SM. Overexpression of KLF13 and FGFR3 in oral cancer cells. Cytogenet Genome Res. 2010 Jun;128(4):192-8. doi: 10.1159/000308303. Epub 2010 Jun 2.

Hinks A, Martin P, Flynn E, Eyre S, Packham J; Childhood Arthritis Prospective Study-CAPS; BSPAR Study Group, Barton A, Worthington J, Thomson W. Subtype specific genetic associations for juvenile idiopathic arthritis: ERAP1 with the enthesitis related arthritis subtype and IL23R with juvenile psoriatic arthritis. Arthritis Res Ther. 2011 Jan 31;13(1):R12. doi: 10.1186/ar3235.

Kanazawa A, Kawamura Y, Sekine A, Iida A, Tsunoda T, Kashiwagi A, Tanaka Y, Babazono T, Matsuda M, Kawai K, Iiizumi T, Fujioka T, Imanishi M, Kaku K, Iwamoto Y, Kawamori R, Kikkawa R, Nakamura Y, Maeda S. Single nucleotide polymorphisms in the gene encoding Krüppel-like factor 7 are associated with type 2 diabetes. Diabetologia. 2005 Jul;48(7):1315-22. Epub 2005 Jun 4.

Karlsson R, Aly M, Clements M, Zheng L, Adolfsson J, Xu J, Grönberg H, Wiklund F. A Population-based Assessment of Germline HOXB13 G84E Mutation and Prostate Cancer Risk. Eur Urol. 2012 Jul 20. [Epub ahead of print]

Mehta AM, Jordanova ES, Corver WE, van Wezel T, Uh HW, Kenter GG, Jan Fleuren G. Single nucleotide polymorphisms in antigen processing machinery component ERAP1 significantly associate with clinical outcome in cervical carcinoma. Genes Chromosomes Cancer. 2009 May;48(5):410-8.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006 Aug;38(8):904-9. Epub 2006 Jul 23.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC. PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics. 2007, 81.

Saveanu L, Carroll O, Lindo V, Del Val M, Lopez D, Lepelletier Y, Greer F, Schomburg L, Fruci D, Niedermann G, van Endert PM. Concerted peptide trimming by human ERAP1 and ERAP2 aminopeptidase complexes in the endoplasmic reticulum. Nat Immunol. 2005 Jul;6(7):689-97. Epub 2005 May 22.

Szczypiorska M, Sánchez A, Bartolomé N, Arteta D, Sanz J, Brito E, Fernández P, Collantes E, Martínez A, Tejedor D, Artieda M, Mulero J. ERAP1 polymorphisms and haplotypes are associated with ankylosing spondylitis susceptibility and functional severity in a Spanish population. Rheumatology (Oxford). 2011 Nov;50(11):1969-75. doi: 10.1093/rheumatology/ker229. Epub 2011 Aug 24.

Scohy S, Gabant P, Van Reeth T, Hertveldt V, Drèze PL, Van Vooren P, Rivière M, Szpirer J, Szpirer C. Identification of KLF13 and KLF14 (SP6), novel members of the SP/XKLF transcription factor family. Genomics. 2000 Nov 15;70(1):93-101.

Thorne H., Willems A.J., Niedermayr E., et al. Decreased prostate cancer-specific survival of men with BRCA2 mutations from multiple breast cancer families. Cancer Prev Res (Phila). 4, 1002-1010 (2011).

Trembath RC Genetic Analysis of Psoriasis Consortium & the Wellcome Trust Case Control Consortium 2, Strange A, Capon F, Spencer CC, Knight J, Weale ME, Allen MH, Barton A, Band G, Bellenguez C, Bergboer JG, Blackwell JM, Bramon E, Bumpstead SJ, Casas JP, Cork MJ, Corvin A, Deloukas P, Dilthey A, Duncanson A, Edkins S, Estivill X, Fitzgerald O, Freeman C, Giardina E, Gray E, Hofer A, Hüffmeier U, Hunt SE, Irvine AD, Jankowski J, Kirby B, Langford C, Lascorz J, Leman J, Leslie S, Mallbris L, Markus HS, Mathew CG, McLean WH, McManus R, Mössner R, Moutsianas L, Naluai AT, Nestle FO, Novelli G, Onoufriadis A, Palmer CN, Perricone C, Pirinen M, Plomin R, Potter SC, Pujol RM, Rautanen A, Riveira-

Munoz E, Ryan AW, Salmhofer W, Samuelsson L, Sawcer SJ, Schalkwijk J, Smith CH, Ståhle M, Su Z, Tazi-Ahnini R, Traupe H, Viswanathan AC, Warren RB, Weger W, Wolk K, Wood N, Worthington J, Young HS, Zeeuwen PL, Hayday A, Burden AD, Griffiths CE, Kere J, Reis A, McVean G, Evans DM, Brown MA, Barker JN, Peltonen L, Donnelly P, Trembath RC. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1.Nat Genet. 2010 Nov;42(11):985-90. doi: 10.1038/ng.694. Epub 2010 Oct 17.

Tryggvadottir L., Vidarsdottir L., Thorgeirsson T., et al. Prostate cancer progression and survival in BRCA2 mutation carriers. J Natl Cancer Inst. 99, 929-935 (2007).

Wu M.C., Lee S., Cai T., et al. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 89, 82-93 (2011).

Wan X, Yang C, Yang Q, Xue H, Tang NL, Yu W.Predictive rule inference for epistatic interaction detection in genome-wide association studies. Bioinformatics. 2010 Jan 1;26(1):30-7. Epub 2009 Oct 30.

Yan J, Parekh VV, Mendez-Fernandez Y, Olivares-Villagómez D, Dragovic S, Hill T, Roopenian DC, Joyce S, Van Kaer L. In vivo role of ER-associated peptidase activity in tailoring peptides for presentation by MHC class Ia and class Ib molecules. J Exp Med. 2006 Mar 20;203(3):647-59. Epub 2006 Feb 27.

Zhang,Y. and Liu,J.S. (2007) Bayesian inference of epistatic interactions in case-control studies. Nat. Genet., 39,1167–1173.